RESEARCH PAPER

Multilingual Noise-Aware Alzheimer's Disease Detection from Speech Using Hyperbolic Multi-View Fusion

Aman Singh[1], Dileep Kumar Yadav[2]
[1]M.Tech Scholar, Dept. of CSE, Faculty of Engineering and Technology. Email: amansingh7317724@gmail.com
[2]Assistant Professor, Dept. of CSE, Faculty of Engineering and Technology. Email: dileep1482@gmail.com

ABSTRACT
Alzheimer's disease (AD) is a progressive neurodegenerative disorder that affects speech and language abilities, making speech analysis a promising non-invasive tool for early screening and monitoring. Most existing speech-based AD detection systems are developed using single-language datasets and controlled recording conditions, which limits their applicability in multilingual and noisy environments. This study presents a multilingual noise-aware speech-based AD detection framework evaluated across English, Chinese, and Spanish. The proposed framework uses pretrained speech encoders as frozen feature extractors, applies controlled noise augmentation at clean, 10 dB, and 15 dB conditions, and introduces HyJS-Fuse, a hyperbolic Jensen–Shannon divergence-based multi-view fusion framework. HyJS-Fuse maps complementary speech representations into a Poincaré ball, fuses them through Möbius addition, and optimizes the fused prediction using a supervised Jensen–Shannon divergence objective with optional branch-consistency regularization. Experimental results show that the proposed framework consistently outperforms single-encoder and concatenation-based baselines across languages and acoustic conditions. The best clean-condition results reach 98.27% accuracy and 96.89% F1 on English, 95.06% accuracy and 93.82% F1 on Chinese, and 91.02% accuracy and 89.24% F1 on Spanish. These findings demonstrate that geometry-aware multi-view fusion is effective for robust multilingual AD detection from speech.

Keywords: Alzheimer's disease; Speech analysis; Multilingual detection; Noise robustness; Hyperbolic fusion; Jensen–Shannon divergence; Pretrained speech encoders.
How to cite this article: Singh A, Yadav DK. Multilingual Noise-Aware Alzheimer's Disease Detection from Speech Using Hyperbolic Multi-View Fusion. Int J Drug Deliv Technol. 2026;16(55s): 823-827. DOI: 10.25258/ijddt.16.55s.83
Source of support: Nil.
Conflict of interest: None.

1. INTRODUCTION

Alzheimer's disease (AD) is a progressive neurodegenerative disorder associated with gradual cognitive decline and substantial clinical burden.[1,2] Early identification of individuals at elevated risk or in early-stage decline is important for timely clinical management and improved long-term outcomes.[3] However, conventional assessment pathways often depend on specialized clinical testing and structured neuropsychological evaluation, which can be resource-intensive and difficult to scale in low-resource or large-screening settings.[4]

Speech-based assessment provides a promising non-invasive alternative because speech can be collected at low cost and can reflect changes in articulation, fluency, prosody, lexical retrieval, and temporal organization. Prior studies have shown that acoustic and linguistic speech markers can support automatic AD detection using machine learning and speech processing methods.[5,6] Nevertheless, most existing systems are developed under monolingual and controlled recording conditions. This limits their generalization to multilingual populations and realistic acoustic environments where background noise, channel variation, and recording artifacts are common.[7]

Another limitation is that many speech-based AD systems rely on either a single feature representation or simple concatenation of multiple features. Such approaches may fail to preserve complementary structure across different pretrained speech encoders. This is especially important in multilingual AD detection, where speech representations may capture different phonetic, prosodic, and language-specific cues. Motivated by these limitations, this study proposes HyJS-Fuse, a hyperbolic Jensen–Shannon divergence-based multi-view fusion framework for multilingual noise-aware AD detection from speech.

The main contributions of this work are:

- We present a multilingual speech-based AD detection framework evaluated across English, Chinese, and Spanish under clean and noise-degraded conditions.
- We investigate multiple pretrained speech encoders as frozen feature extractors to capture complementary acoustic and linguistic information.
- We propose HyJS-Fuse, a hyperbolic multi-view fusion framework that combines

encoder representations using Möbius addition and Jensen–Shannon divergence-based learning.

- We show that the proposed framework consistently improves over single-encoder and concatenation baselines across languages and acoustic conditions.

2. RELATED WORK

Speech-based AD detection has been widely studied as a non-invasive approach for cognitive screening. Earlier work demonstrated that automatic speech analysis can identify markers associated with predementia and AD using acoustic, lexical, and temporal features.[4] Low-resource and challenge-based studies further showed that speech and language technologies can support AD detection and assessment, while also highlighting the sensitivity of model performance to dataset design and evaluation protocol.[5,6]

Noise robustness and feature stability remain important challenges for clinical speech systems. Real-world recordings may include background noise, microphone variability, and environmental distortions. Prior work on noise-robust speech processing has shown that acoustic degradation can significantly affect recognition and downstream classification performance.[7] Similarly, heterogeneous speech feature design and careful feature selection can improve model stability in limited-data settings.[8]

Recent pretrained speech encoders such as XLS-R, wav2vec 2.0, Whisper, MMS, and TRILLsson provide transferable representations learned from large-scale speech data. These encoders capture different aspects of speech, including phonetic, acoustic, multilingual, and semantic information. However, simply concatenating encoder representations may not fully exploit their complementary structure. This motivates the use of geometry-aware fusion, where different representation streams can be integrated in a structured latent space.

3. MATERIALS AND METHODS

3.1. Pretrained Speech Encoders

To capture complementary speech characteristics relevant to AD detection, we use multiple pretrained speech representation models as frozen feature extractors. Specifically, we consider TRILLsson, XLS-R, wav2vec 2.0, Whisper, and MMS. These encoders differ in their training objectives, multilingual exposure, and representational focus. All encoders are kept frozen during training to ensure stable comparison across languages and acoustic conditions.

3.2. Dataset Description

The study uses a multilingual speech dataset comprising English, Chinese, and Spanish. These languages were selected to evaluate cross-language robustness across distinct phonetic, prosodic, and structural characteristics.

For English, we use the Pitt Corpus from DementiaBank.[9] The Cookie Theft picture description task is used, following common practice in dementia speech analysis. The English subset contains 243 healthy control (HC) samples and 309 AD samples. For Spanish, we use the Ivanova corpus from DementiaBank,[10] which contains standardized reading speech recordings from older adults. The Spanish subset includes 196 HC samples and 74 AD samples. For Chinese, we use the NCMMSE dataset released through NCMMSC Alzheimer's recognition resources.[11] The Chinese subset contains 108 HC samples and 79 AD samples.

Table 1: Summary of multilingual AD speech datasets.

Language	Dataset	HC Samples	AD Samples
English	Pitt Corpus	243	309
Chinese	NCMMSE	108	79
Spanish	Ivanova Corpus	196	74

3.3. Noise Augmentation Strategy

To evaluate robustness under realistic acoustic conditions, controlled noise augmentation is applied by adding background noise to clean speech signals at predefined signal-to-noise ratio (SNR) levels. Let $x(t)$ denote a clean speech waveform and $n(t)$ denote a noise waveform segment of the same length. The augmented signal is defined as:

$$y(t) = x(t) + \alpha\, n(t) \quad (1)$$

where α is selected to satisfy the target SNR:

$$SNR_{\mathrm{dB}} = 10 \log_{10} \left(\|x\|_2^2 / \|\alpha n\|_2^2 \right) \quad (2)$$

Solving for α gives:

$$\alpha = \sqrt{\left(\|x\|_2^2 / \left(\|n\|_2^2 \cdot 10^{\mathrm{SNR\_dB}/10} \right) \right)} \quad (3)$$

Noise augmentation is applied only after speaker-independent train–test partitioning to avoid data leakage. The same augmentation protocol is used across English, Chinese, and Spanish.

3.4. Proposed HyJS-Fuse Framework

We propose HyJS-Fuse, a hyperbolic Jensen–Shannon divergence-based multi-view fusion framework. Fig. 1 shows the overall model architecture.

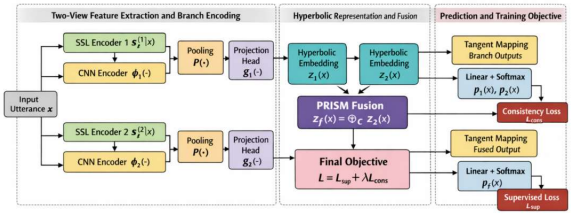

Figure 1: Overview of the proposed HyJS-Fuse framework. Two complementary speech representation views are extracted using frozen SSL encoders, processed through branch-specific CNN encoders and projection heads, mapped into hyperbolic space, and fused through Möbius addition. The fused and branch-wise embeddings are mapped to the tangent space for probabilistic prediction and optimized using supervised and branch-consistency objectives.

Let $D = \{(x_n, y_n)\}^N_{n=1}$ denote the training dataset, where x_n is an input utterance and y_n is the corresponding class label. For each utterance x, two complementary SSL feature views are extracted and denoted as $s^{(1)}(x)$ and $s^{(2)}(x)$. Each view is passed through a branch-specific CNN encoder $\phi_b(\cdot)$ and pooling operator $P(\cdot)$ to obtain:

$$h_b(x) = P(\phi_b(s^{(b)}(x))) \quad (4)$$

where $b \in \{1, 2\}$. A projection head $g_b(\cdot)$ then maps each branch representation into a Euclidean latent vector:

$$u_b(x) = g_b(h_b(x)) \quad (5)$$

Each Euclidean latent vector is mapped to a Poincaré ball $B^d_c = \{z \in \mathbb{R}^d : c\|z\|^2 < 1\}$ using the exponential map at the origin. This gives branch-specific hyperbolic embeddings $z_1(x)$ and $z_2(x)$. The two embeddings are fused using Möbius addition:

$$z_f(x) = z_1(x) \oplus_c z_2(x) \quad (6)$$

For two points $p, q \in B^d_c$, Möbius addition is defined as:

$$p \oplus_c q = [(1 + 2c\langle p, q\rangle + c\|q\|^2)p + (1 - c\|p\|^2)q] / [1 + 2c\langle p, q\rangle + c^2\|p\|^2\|q\|^2] \quad (7)$$

The fused embedding and branch embeddings are mapped back to the tangent space using the logarithmic map. Linear classifiers then produce branch-wise and fused logits. These logits are converted into predictive probability distributions using softmax. The Jensen–Shannon divergence between two distributions p and q is:

$$D_{JS}(p, q) = H((p + q)/2) - ½ H(p) - ½ H(q) \quad (8)$$

where $H(\cdot)$ denotes Shannon entropy. The supervised loss is:

$$L_{sup}(x, y) = D_{JS}(e(y), p_f(\tilde{x})) \quad (9)$$

where $e(y)$ is the one-hot target distribution and $p_f(\tilde{x})$ is the fused prediction from an augmented or perturbed input. The final objective combines supervised loss with branch-consistency regularization:

$$L(x, y) = L_{sup}(x, y) + \lambda L_{cons}(x) \quad (10)$$

where λ controls the consistency term.

3.5. Classification and Evaluation Protocol

All experiments use speaker-independent splits to ensure that no subject appears in both training and testing sets. Performance is evaluated separately for each language under clean, 10 dB, and 15 dB conditions. Accuracy and F1-score are reported as the main evaluation metrics. We compare three settings: single-encoder features, concatenation-based pairwise fusion, and the proposed HyJS-Fuse framework.

4. RESULTS AND DISCUSSION

To keep the main manuscript concise, we report compact summaries of the strongest results across single-encoder, concatenation, and HyJS-Fuse settings. Complete model-pair results can be included as supplementary material.

4.1. Single-Encoder Results

Table 2 summarizes the strongest single-encoder results. XLS-R achieves the strongest clean-condition F1 on English and Chinese, while TRILLsson is also competitive. MMS and wav2vec 2.0 are generally weaker, indicating that multilingual and semantically rich encoders provide more useful AD-related cues.

Table 2: Best single-encoder results across languages.

Language	Best Encoder	Clean F1	10 dB F1	15 dB F1
English	XLS-R	87.02	85.64	85.23
Chinese	XLS-R	83.45	82.38	81.26
Spanish	XLS-R	80.04	78.93	77.29

4.2. Concatenation Baseline

Pairwise concatenation improves over single encoders in most settings, especially for pairs involving PaSST, XLS-R, and Whisper. However, the improvements are moderate and degrade under stronger noise. This suggests that direct feature joining does not fully exploit the complementary structure across encoder views.

Table 3: Best concatenation-based pairwise fusion results.

Language	Best Pair	Clean F1	10 dB F1	15 dB F1
English	PaSST + wav2vec 2.0	91.26	87.92	86.79
Chinese	XLS-R + MMS	88.21	87.13	85.37
Spanish	PaSST + XLS-R	83.51	81.76	79.64

4.3. Proposed HyJS-Fuse Results

Table 4 summarizes the strongest HyJS-Fuse results. The proposed framework consistently improves over both single-encoder and concatenation baselines across all three languages. On English, PaSST + XLS-R achieves 98.27 accuracy and 96.89 F1 in the clean condition. On Chinese, XLS-R + wav2vec 2.0 achieves 95.06 accuracy and 93.82 F1. On Spanish, XLS-R + Whisper achieves 91.02 accuracy and 89.24 F1. These results show that hyperbolic fusion preserves cross-view information more effectively than simple concatenation.

Table 4: Best HyJS-Fuse results across languages.

Language	Best Pair	Clean F1	10 dB F1	15 dB F1
English	PaSST + XLS-R	96.89	95.67	94.17
Chinese	XLS-R + wav2vec 2.0	93.82	92.11	90.56
Spanish	XLS-R + Whisper	89.24	87.39	86.53

4.4. Discussion

The experimental findings support three main conclusions. First, multilingual AD detection benefits from representation diversity, especially when complementary pretrained encoders are combined. Second, direct concatenation provides useful gains over single encoders but remains limited under noise. Third, HyJS-Fuse provides the most consistent improvements across all languages and acoustic conditions, suggesting that geometry-aware fusion is effective for preserving complementary cross-view structure.

Spanish remains the most challenging dataset, likely due to dataset size, task structure, or acoustic variability. However, the proposed method still improves substantially over the baselines. These results indicate that hyperbolic multi-view fusion is a promising direction for multilingual cognitive health assessment from speech.

5. CONCLUSION

This study presented a multilingual noise-aware speech-based Alzheimer's disease detection framework evaluated across English, Chinese, and Spanish. The proposed HyJS-Fuse framework combines complementary pretrained speech representations in hyperbolic space and optimizes the final prediction using a Jensen–Shannon divergence-based objective. Experimental results show that HyJS-Fuse consistently outperforms single-encoder and concatenation-based baselines across clean, 10 dB, and 15 dB conditions. These findings demonstrate that geometry-aware multi-view fusion is an effective strategy for robust multilingual AD detection from speech. Future work will investigate broader multilingual settings, cross-dataset generalization, component-wise ablation studies, and alternative divergence-based objectives for improved stability and interpretability.

ACKNOWLEDGEMENT

The authors are thankful to the Faculty of Engineering and Technology (UNSIET), Veer Bahadur Singh Purvanchal University, Jaunpur, Uttar Pradesh, India, for providing academic support.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

AUTHOR CONTRIBUTIONS

Aman Singh contributed to conceptualization, methodology, experimentation, result analysis, and manuscript writing. Dileep Kumar Yadav contributed to supervision, validation, critical review, and manuscript revision. Both authors reviewed and approved the final manuscript.

References

[1] Tapan Behl, Gagandeep Kaur, Aayush Sehgal, Shaveta Bhardwaj, Sukhbir Singh, Camelia Buhas, Claudia Judea-Pusta, Diana Uivarosan, Mihai Alexandru Munteanu, and Simona Bungau. Multifaceted role of matrix metalloproteinases in neurodegenerative diseases: Pathophysiological and therapeutic perspectives. *International Journal of Molecular Sciences*, 22(3):1413, 2021.

[2] Tanima Bhattacharya, Giselle Amanda Borges e Soares, Hitesh Chopra, Md Mominur Rahman, Ziaul Hasan, Shasank S Swain, and Simona Cavalu. Applications of phytonanotechnology for the treatment of neurodegenerative disorders. *Materials*, 15(3):804, 2022.

[3] Karen Ritchie, Isabelle Carriere, Li Su, John T O'Brien, Simon Lovestone, Katie Wells, and Craig W Ritchie. The midlife cognitive profiles of adults at high risk of late-onset alzheimer's disease: The prevent study. *Alzheimer's & Dementia*, 13(10):1089–1097, 2017.

[4] Alexandra König, Aharon Satt, Alexander Sorin, Ron Hoory, Orith Toledo-Ronen, Alexandre Derreumaux, Valeria Manera, Frans Verhey, Pauline Aalten, Phillipe H Robert, et al. Automatic speech analysis for the assessment of patients with predementia and alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 1(1):112–124, 2015.

[5] Raghavendra Pappagari, Jaejin Cho, Sonal Joshi, Laureano Moro-Velazquez, Piotr Żelasko, Jesus Villalba, and Najim Dehak. Automatic detection and assessment of alzheimer disease using speech and language technologies in low-resource scenarios. In *Interspeech*, volume 2021, pages 3825–3829, 2021.

[6] Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. Detecting Cognitive Decline Using Speech Only: The ADReSSo Challenge. In *Interspeech 2021*, pages 3780–3784, 2021. doi: 10.21437/Interspeech.2021-1220.

[7] Puneet Bawa and Virender Kadyan. Noise robust in-domain children speech enhancement for automatic punjabi recognition system under mismatched conditions. *Applied Acoustics*, 175:107810, 2021.

[8] Virender Kadyan, Archana Mantri, and RK Aggarwal. A heterogeneous speech feature vectors generation approach with hybrid hmm classifiers. *International Journal of Speech Technology*, 20(4):761–769, 2017.

[9] TalkBank / DementiaBank. Dementiabank: Pitt corpus (english). https://talkbank.org/dementia/access/English/Pitt.html. Accessed 2026-02-04.

[10] TalkBank / DementiaBank. Dementiabank: Ivanova corpus (spanish). https://talkbank.org/dementia/access/Spanish/Ivanova.html. Accessed 2026-02-04.

[11] National Conference on Man–Machine Speech Communication (NCMMSC). Ncmmsc 2021 speech corpus / ncmmse (chinese alzheimer's disease recognition resources). http://www.ncmmsc2021.org/. Accessed 2026-02-04.